Trends in Food Science & Technology 99 (2020) 181–193

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: www.elsevier.com/locate/tifs

On the pros and cons of Bayesian kinetic modeling in food science

M.A.J.S. van Boekel

Food Quality & Design Group, Wageningen University & Research, P.O. Box 8129, 6700, EV Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Kinetics
Bayesian statistics
Least-squares regression
Modeling
Brms
Stan

ABSTRACT

Background: Kinetics is an important part of food science and statistics is a necessary key element in modeling. Ordinary least-squares (OLS) regression is mostly used to obtain parameter estimates and their uncertainties; this is done within the frequentist framework.

Scope and approach: This article introduces Bayesian statistics as an alternative to OLS. The background of Bayesian statistics is briefly explained, emphasizing the difference with the frequentist approach. Basically, frequentists go for the probability of data given a hypothesis, resulting in point estimates, while Bayesians go for the probability of a hypothesis given the data, resulting in probability distributions for parameters. This study shows how to apply the Bayesian approach to kinetic problems using freely available R packages. To focus on the Bayesian approach, the kinetic problem presented is a trivial zero-order reaction concerning the formation of furan in a soy sauce.

Key findings and conclusions: The main result is numerical and graphical output showing probability distributions of parameters. Interpretation of regression results is shown leading to the conclusion that the Bayesian approach yields a more intuitive result with richer information than the conventional OLS approach. The pros and cons of the Bayesian approach are highlighted, the major pro being the intuitive and informative result and the major con that one has to learn and apply a programming language like R or Python. The Bayesian approach is very general and the outline shown here can be applied easily to much more complicated kinetic models.

1. Introduction

The title of this paper reflects two key phrases: "kinetic modeling" on the one hand, and "Bayesian" on the other. As an introduction to these concepts: suppose one is interested in the fate of, say, vitamin C in a food, hypothesizing that its concentration changes because of a chemical reaction (e.g., oxidation) and as a first idea the concentration of vitamin C is assumed to decrease linearly in time; in doing so, a kinetic model is proposed. Then, by measuring vitamin C, it appears not to decrease linearly but, even though there is variation in the data, it seems to decrease more exponentially. Knowledge is updated accordingly and an exponential model is proposed. Working and thinking like this, one applies, perhaps unknowingly, Bayesian reasoning: an uncertain, prior idea is combined with uncertain data and as a result knowledge is updated and uncertainty decreases. Bayesian statistics is the formal way to apply such reasoning, while kinetics is the formal way to model changes over time. Even though describing this way of thinking seems like an open door, it is not the way classical statistics works. The usual statistical approach is the frequentist approach based on sampling theory. An important difference is how model parameters are dealt with, like rate constants and activation energies in kinetics. Such parameters cannot be observed or measured but they can be inferred from data. As will be explained below, Bayes' theorem is very useful in that respect. A fundamental difference in considering parameters will appear between the classical frequentist approach and the Bayesian approach. The main purpose of this paper is to explain the alternative approach of Bayesian reasoning and how that can be applied to kinetic modeling.

Kinetics is an indispensable part of food science and has many applications in product and process design, and it offers, of course, also a scientific tool to understand reactions in foods. Kinetic models are, therefore, frequently discussed in food science journals, including the statistical aspects of analysis of data, e.g., Goula, Prokopiou, and Stoforos (2018), Grainger, Owens, Manley-Harris, Lane, and Field (2017), Zhang, Chen, Boom, and Schutyser (2017), Rial-Otero, Simal-Gándara, Pose-Juan, López-Fernández, and Yáñez (2018), to cite just a few. Often, excel is used as the software tool to analyse kinetic data (e.g., Hites (2017)), even though it is well recognized that it lacks some necessary statistical tools (De Levie, 2012). The nonlinear Solver routine in Excel is not always reliable (Tellinghuisen, 2015) and does not even supply any uncertainty in parameters; fortunately, some of these failures can be remedied by using additional macro's, for instance as developed by de

E-mail address: tiny.vanboekel@wur.nl.

https://doi.org/10.1016/j.tifs.2020.02.027
Received 27 May 2019; Received in revised form 5 August 2019; Accepted 25 February 2020
Available online 05 March 2020

Levie. This concerns specifically the macro's LS1 (linear regression with intercept), Solveraid (calculates parameter uncertainties after usage of the Solver for nonlinear regression) and Propagation (estimates how uncertainties in parameters are propagated in subsequent calculations) in the so-called Macrobundle, see website de Levie (http://www.bowdoin.edu/~rdelevie/excellaneous/#macrobundlecontents). Another example of an add-in to Excel is given by Halpern, Frye, and Marzzacco (2018). Of course, also many other software programmes than Excel are possible, commercially or freely available. Almost all of the statistical software tools employ what is called ordinary least squares (OLS) regression, be it linear or nonlinear. Appreciating the statistical aspects of kinetic modeling is not always easy and would deserve some more attention in publications in the view of the present author. The common (and dominant) way of statistical modeling is based on the already mentioned frequentist approach, the branch of statistics that is usually taught at high schools and universities. Not too many people active in kinetics seem to apply Bayesian statistics as another approach that could be used in kinetic modeling. There are some exceptions in chemical engineering. A Bayesian approach was developed in the 1960s (Box & Draper, 1965) to handle multireponse problems in chemical engineering, see also Stewart, Caracotsios, and Sørensen (1992) and Stewart and Caracotsios (2008). This approach is the basis for a software programme called Athena Visual Studio developed especially for chemical engineering problems (http://www.athenavisual.com) that contains Bayesian options. Apart from that, there are only few references to be found that use Bayesian methods in kinetics, e.g., Galagali and Marzouk (2015). However, the Bayesian approach is used a lot in other branches of science, like ecology, e.g., Hobbs and Hooten (2015), Korner-Nievergelt et al. (2016), and in the medical sciences e.g, Zwinderman (2018); interestingly, it is much more popular in an area close to kinetic modeling, namely in pharmacokinetic modeling, e.g., Krauss, Tappe, Schuppert, Kuepfer, and Goerlitz (2015), Carreno, Lomaestro, Tietjan, and Lodise (2017).

It is proposed here that the Bayesian approach could be an attractive alternative to the traditional approach because it is more intuitive as to how people think regarding statistics. Until recently, there was a good reason to stay with the frequentist way because the Bayesian way was computationally demanding, or even impossible. However, that has changed the past two decades, for two reasons. First, the software to handle statistics in the Bayesian way has enormously improved and can now be used routinely (as will be shown here). Second, the software to do that is freely accessible via the software package R, so basically everyone who has a computer and access to internet can use it. Over the last decade, R has become very popular, and is rapidly turning into the standard for statistical modeling as well as numerous other applications. Next to that R is completely free and available for every computer platform, it is supported by very active user groups. As a result, many dedicated packages have been, and are being developed, freely available, well documented and, most importantly, supported by a large community. Chances are that a particular problem has already, partly perhaps, been tackled by someone else, and if not, there is always someone prepared to look at the problem at hand and come up with suggestions for a solution. "Version control systems" like Git, Github, Bitbucket are directly linked to R and RStudio, thereby enabling community of workers to communicate and interact, thus contributing to open science (Bryan, 2018). There are many specific websites for getting help, as an internet search will quickly show. Next to R, other software programs like Matlab, Mathematica, MathCad, can be used as well, of course. Also these programs have user groups, good documentation, are powerful and user-friendly, but they are not for free and therefore not accessible to everyone like R is. The purpose of this publication is twofold. First, to show the advantage of the Bayesian approach, which is suggested to be more intuitive than frequentist statistics, and conceptually simple. Second, the purpose is to show how Bayesian statistics can be used easily in kinetic modeling and how more information can be extracted from data than with the frequentist approach. The investment to be made is that one needs to learn the language of R (or a similar language like Python, for example). While that is indeed an investment, it is one that pays off in many ways, an important one being that companies and universities alike are increasingly using R. Since R is open access, codes are freely published and shared, which is a great help. The outline for this paper is as follows. First, a brief recapitulation of regression is given, followed by an introduction to the Bayesian approach to regression. This will be very limited because there are many excellent tutorials to be found on internet, universities offer courses, papers and books are available for more details. Some introductory papers can be found here (though not for food science problems): Muth, Oravecz, and Gabry (2018), Baldwin and Larson (2017), Eguchi (2008). Three books are definitely worth mentioning for those interested in more background reading. An essential, and absolutely ground breaking one for novices in modeling and statistics is "Statistical Rethinking"" by McElreath (2016), another recently published and very instructive book is "A student guide to Bayesian Statistics" by Lambert (2018), and the third book worth mentioning is "Doing Bayesian Data Analysis"" by Kruschke (2015). These authors focus on the concepts with a minimum of mathematical treatment and give many examples in R.

2. Theoretical background

2.1. Statistical modeling: regression revisited

Kinetic parameters need, ultimately, to be derived from experimental measurements. Experiments yield variable (and therefore uncertain) results. Consequently, kinetic parameters estimated from these experiments are also uncertain, and this uncertainty needs to be characterized, which is why we need statistics. The statistical technique to estimate parameters from experimental data is called regression. Let us briefly recapitulate what we are actually doing then. Recall that the purpose of kinetics is, simply said, to investigate how the rate of a reaction changes over time and with temperature. Relevant parameters in that respect are: the order of a reaction, rate constants and activation parameters (activation energy, activation entropy and enthalpy). Based on the law of mass action, the rate of a reaction of a component (dc/dt) is measured via the concentration of the reactant (c) as a function of time (t):

$$\frac{dc}{dt} = k_r \cdot c^{n_t} \tag{1}$$

properly called the general rate law equation. The two parameters in this equation are the proportionality constant k_r (in kinetics called the rate constant) and the order n_t. In many cases in food science literature, authors assume a certain value for n_t, mostly 0, 1, or 2. For instance, integration of equation (1) for a zero-order reaction, i.e., $n_t = 0$, for a reaction where a compound is formed yields a linear relation between concentration and time:

$$c = c_0 + k_r \cdot t \tag{2}$$

in which c_0 is the initial concentration. (For a degradation reaction the + sign becomes a - sign.) If $n_t = 1$, a first-order model arises which upon integration of equation (1) shows that the change in concentration depends exponentially on time, in this case for a degradation reaction:

$$c = c_0 \cdot \exp(-k_r \cdot t) \tag{3}$$

Equation (1) can also be integrated as a whole, which yields:

$$c = \left(c_0^{1-n_t} + (n_t - 1)k_r \cdot t\right)^{\frac{1}{1-n_t}} \tag{4}$$

By fitting this equation to data, the order n_t can be estimated from experimental data next to c_0 and k_r. The order is usually found to be between 0 and 3; note that it can also be fractional (see Van Boekel (2008) for more details).

The rate constant and its dependence on temperature (not discussed

here to keep things simple), and if needed the order n_t, are to be estimated from experimental data. With such knowledge we can predict the rate of a reaction according to the kinetic model proposed. The proposed kinetic model is called a deterministic model because it completely determines the outcome if we would know the parameters exactly. To find these parameters we have to confront the model with reality, in food studies usually data consisting of chemical concentrations or physical measurements (called the response variable) as a function of time t and temperature T, and occasionally pressure P (called the independent, predictor variable or covariate).

However, as mentioned, experimental data (commonly symbolized by y_i) are always variable (and hence uncertain) and so, we need a statistical model that describes this variability. It is well known that experimental variation resulting from chemical and physical measurements is usually well described by a normal distribution. This follows from the central limit theorem stating that the outcome of many small, unpredictable variations (as happens during measurements) always lead to a normal distribution. A normal distribution is completely characterized by its mean μ and variance σ^2:

$$p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad (5)$$

This equation should be read as: the probability of observing certain values of experimental data y can be predicted if we know the mean and variance of the process that generates the experimental data y. So, in kinetics the goal is to estimate this mean and standard deviation for each measurement at each measured point x_i: see Fig. 1.

We can thus connect a measured value (for instance, the concentration of a chemical measured at a certain value x) to a variable representing a central value (like the mean μ) according to a model, while the experimental variation is described by the parameter σ^2 controlling the dispersion of a distribution. This makes it a statistical model, which is for kinetics usually (but not necessarily) a Gaussian distribution.

Writing this reasoning in mathematical/statistical language, with a zero-order formation reaction (equation (2)) as example leads to the following statistical model for the variation in the experimental data:

$$y_i \sim \mathcal{N}(\mu_i, \sigma) \quad (6)$$

The expectation value $E(y_i)$, or the mean μ_i, is coupled to the predictor variable time t_i as specified by the deterministic kinetic model, in this example the zero-order model:

$$\mu_i = c_0 + k_r \cdot t_i \quad (7)$$

Equation (6) should be read as: the variation in each experimental data point y_i (with $i = 1...n$ datapoints) is assumed to be normally distributed with mean μ_i and a constant standard deviation σ (so, σ is not supposed to depend on t_i, see Fig. 1). μ_i is according to the proposed model completely determined by c_0 and k_r; the connection between the measured y_i and the model values μ_i is thus via equations (6) and (7). Note that μ_i itself does not need to be estimated because it can be calculated directly from the two to be estimated parameters via a deterministic relationship as expressed in Eqn. (7); uncertainty in μ_i results directly from propagation of uncertainty in the parameters. The case that the standard deviation is constant over the whole range of measurements is called homoscedastic in statistical jargon, if σ varies with the predictor variable (time t_i in this case) the data are called heteroscedastic. To handle heteroscedastic situations in the frequentist framework, see for instance Chapter 7 in Van Boekel (2008), for the Bayesian framework, see Chapter 14 in McElreath (2016).

Armed with this knowledge we can now make the step towards regression in the Bayesian and frequentist framework. First we describe this in general terms and then we focus on kinetic analysis.

2.2. The Bayesian and the frequentist approach compared

At the basis of Bayesian statistics is Bayes' theorem, hence the name (Bayes was an English clergyman who derived his theorem in the 18th century). Before explaining that, let us very briefly recall what we try to do in science. When studying a particular problem, we may build a hypothesis about the phenomenon studied (e.g., that the kinetics can be described by a zero- or first-order reaction), and based on that, we may design experiments to test the hypothesis, and if necessary, adjust the hypothesis if the outcome suggests that the original idea was not correct. Data are obtained from experiments (or by doing observations) and based on such samples an inference is made. In the frequentist framework, we try to generalize from the sample towards the whole population. The parameters that characterize this population (e.g., the mean and the standard deviation, or the rate constant in a kinetic equation) are assumed to be fixed, whereas the data are considered random, i.e., they contain unexplicable variation. In that view, a dataset represents one possible outcome from the many that could have been collected. Inference is then made by referring to a theoretical, infinite number of repetitions (hence the name frequentist). The approach leads to hypothesis testing and p-values, based upon long-run frequencies of repeated sampling from a population. It is about probabilities of obtaining the data as they were obtained and NOT about parameters and hypotheses, and this is, unfortunately, not always realized by researchers. A p-value tells actually more about the data: what is the probability that the sample size is large enough to detect an effect or a relationship. Concepts such as significant or insignificant p-values and confidence intervals are more and more criticized (e.g., Nuzzo (2014); Baker (2016)), not because it is wrong but because many people do not really know how to deal with it and interpret them in a wrong way. This

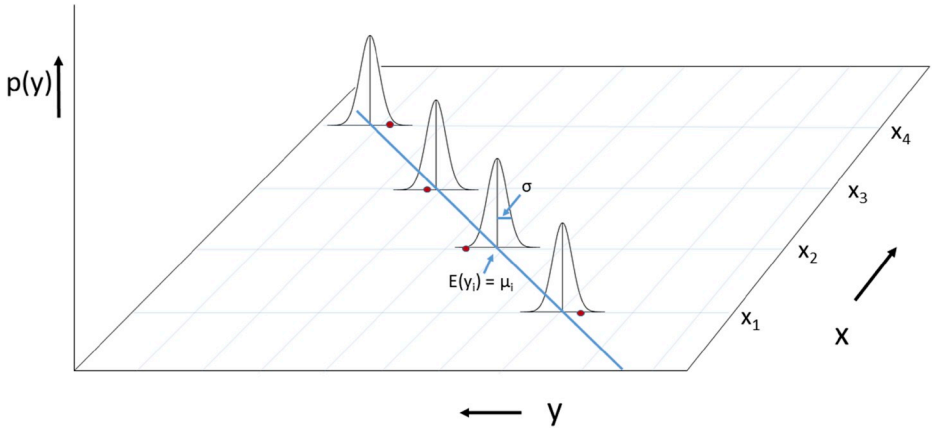

Fig. 1. Schematic representation of a normal probability density $p(y)$ of experimental measurements y_i (indicated by dots) at settings x_i. The line represents a linear model $y = a + b \cdot x$. The expected value at x_i is $E(y_i) = \mu_i$ according to the model, σ is the (constant) standard deviation of the experimental values y_i.

is a debate in its own right that we will not go into here, the purpose is to show that there is an alternative (in which p-values and significance tests are not needed). The focus in the Bayesian perspective is much more on model fit, data visualization, uncertainty and prediction (Baldwin & Larson, 2017). So, what is the difference? In the Bayesian framework, data are considered fixed (fixed does not mean that variation is not acknowledged, it means that we take the data as they are *once they are obtained*). In contrast to the frequentist approach, parameters are being considered random in the sense that we are not sure about their real value and that we can express that uncertainty about parameters in a probabilistic way. In the frequentist framework, however, parameters cannot be variable; they are considered to have a fixed, though unknown, value; and this exact value cannot be estimated precisely because the data show variation. Only if we would be able to do an infinite number of experiments, the exact value would become known to us, but that is obviously not realistic. In the Bayesian framework, the idea is to learn from data to update knowledge about the initial hypothesis and parameters, in other words to decrease uncertainty. Probability is used here to express this uncertainty, as a sort of degree of belief, rather than using probability as a long-run frequency. To summarize, **the frequentist approach is to focus on the probability of data that are obtained if a (null) hypothesis is true, p (data|hypothesis), while Bayesians go for the probability of a hypothesis given the data, p(hypothesis|data)**. To infer something about a hypothesis, we need to go back from an effect that is measured (the data) to the cause that actually produced those data (for which we had an initial hypothesis). This is where Bayes' theorem comes in. This theorem is formally derived from probability theory (derivation is not shown here but can be found in many textbooks, e.g., Lambert (2018); McElreath (2016); Kruschke (2015)) and it links cause and effect such that the cause can be inferred from a measured effect. This is all done in terms of probability distributions that express the uncertainty in a quantitative way, so the outcome is not a point estimate but a probability distribution. This is an important difference with the frequentist approach that only leads to point estimates (in the search for that unknown, fixed parameter value). The confidence intervals that are derived in the frequentist framework about point estimates are NOT probability statements about the parameters (because these have no variation by definition), they are stating the confidence that one can have to obtain that parameter in that specified interval if the experiment would be repeated over and over again. Most people, however, tend to interpret frequentist confidence intervals in the bayesian way, namely as the probability that the parameter is actually in that interval. It should be clear that this is NOT allowed in the frequentist context. Another important difference is that in Bayesian statistics the researcher is forced to state his/her prior opinion about the hypothesis/model/parameter that he/she assumes before the data is analyzed. That is needed to be able to apply Bayes'theorem, which can be formulated as:

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})\cdot p(\text{hypothesis})}{p(\text{data})} \tag{8}$$

The concise mathematical formulation of this is:

$$p(\theta|y) = \frac{p(y|\theta)\cdot p(\theta)}{p(y)} \tag{9}$$

The notation $p(\theta|y)$ should be read as: the probability of θ given the data y and $p(y|\theta)$ as the probability of y given θ (these are so-called conditional probabilities). The left hand side of this equation $p(\theta|y)$ represents the so-called posterior distribution, which is the joint probability distribution of all parameters under consideration, reflecting our knowledge (or ignorance) about the proposed hypothesis, model or parameter (all symbolized by θ) conditional on the data y we have obtained. The numerator on the right hand side is the product of the so-called likelihood $p(y|\theta)$ and the prior $p(\theta)$. The *likelihood* describes the data generating process, given the hypothesis θ, and reflects the probability of observing the data as they have been found, under the assumption that the model is valid. The likelihood is only a proper probability density function (i.e., is positive and sums/integrates to one) when considered as a function of the data with the parameters fixed. However, it is not a true probability density distribution (does not sum to one) when the parameters are considered variable. That is why it is called the likelihood function, expressing the likelihood of obtaining such data under the specified model. For that reason, it is sometimes expressed as $l(y|\theta)$. The *prior* reflects our initial knowledge (or ignorance) about the hypothesis/model/parameter θ before the data are analyzed. Finally, the denominator in Bayes theorem is the marginal distribution of the data, also called the marginal likelihood, (marginal means averaging over all the parameters, the parameters are "integrated out" by considering all possible values of the parameters; the term marginal refers to the margin of a table where the result of the averaging can be shown). In other words, it is the likelihood averaged over the parameters and weighted by their prior probabilities, it is in the end just a number (once the data are obtained), and acts as a normalizing constant by scaling the numerator so that the posterior becomes a real probability density distribution (integrates or sums to one).

Before the advent of computers it was a daunting, and mostly impossible task, to calculate this denominator for all but very simple models. The good news is that we need not to worry anymore about how to calculate the denominator because of the software that allows to numerically approximate the posterior. The denominator does not depend on parameters (which, remember, are integrated out), and it therefore does not carry information about the most probable values of θ. Consequently, it can be left out in the numerical approximation and the basic equation for Bayes' theorem, at least from a conceptual point of view, becomes:

$$p(\theta|y) \propto l(y|\theta) \times p(\theta) \tag{10}$$

(the ∝ sign should be read as: is proportional to). Basically, this result shows that, apart from a scaling factor, the posterior distribution is in fact a combination (weighted average) of the information in the prior (our initial hypothesis θ) and in the data y (expressed in the likelihood) with Bayes theorem being the mathematical machinery to revert quantitatively from knowledge about the data y (which can be observed and expressed in the likelihood function), to the cause of the data as explained by the hypothesis/model/parameter (which can **NOT** be observed, but can be inferred by applying Bayes' theorem).

How do we choose this likelihood and prior? Starting with the likelihood, for kinetic data this can usually be expressed as in equation (6) where it is postulated that the variation in the data can be modelled through the normal distribution. It must be realized, however, that it is an assumption that can and should be checked, e.g., via normal probability plots, or tests like the Shapiro-Wilk test. To put things in perspective, the normal distribution is not always appropriate, for binary outcomes a binomial distribution for the likelihood could be chosen, and for counting problems the Poisson distribution. But for kinetic problems based on chemical/physical measurements a normal distribution will do just fine; alternatively, a close relative of the normal distribution, the t-distribution, could be chosen because it is more robust to outliers (Kruschke, 2015). As for the prior, this is the most debated part of the Bayesian approach because it can have an impact on the resulting posterior, and the debate is then on whether or not a researcher is allowed to subjectively influence the outcome by choosing a prior. As is convincingly shown by many authors (e.g., McElreath (2016); Kruschke (2015); Lambert (2018)), this argument can not be maintained to dismiss the Bayesian approach because subjectivism is present in any scientific endeavour and the nice thing about the Bayesian approach is that one has to specify the prior explicitly, which puts it out in the open and makes it debatable, as it should be in science. Fig. 2 gives some examples of possible priors.

Conceptually, the effect of the prior is as follows. If one is very certain about a model or a parameter (for instance based on literature,

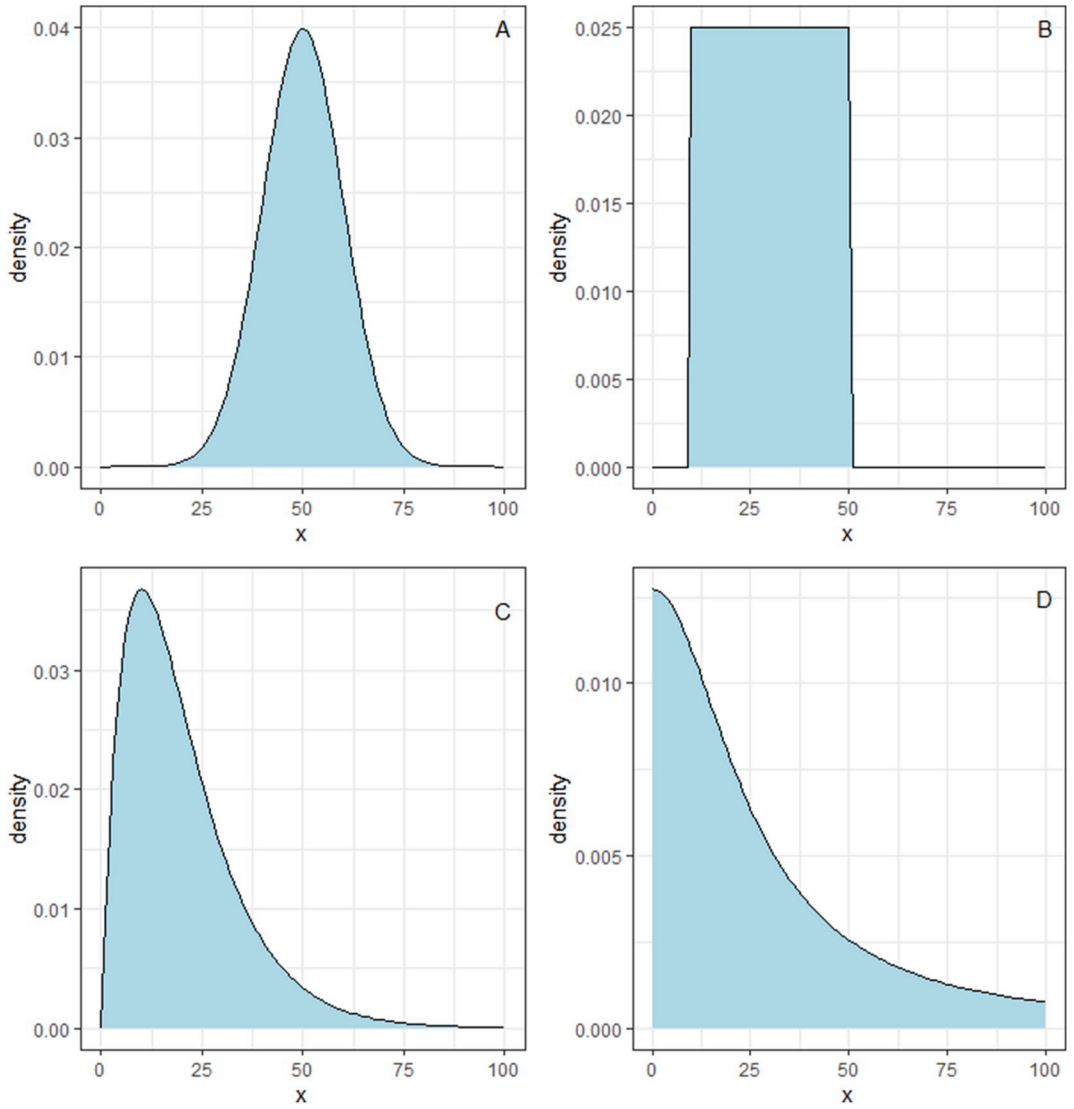

Fig. 2. Examples of possible prior distributions for parameters. A normal distribution with mean = 50 and standard deviation = 10 (A), a uniform distribution between x = 10 and x = 50 (B), a gamma distribution with location parameter = 2 and scale parameter = 0.1 (C), a half-cauchy distribution with location parameter = 0 and scale parameter = 25 (D).

or from earlier experiments), one can propose a strong prior with a narrow distribution, which is then called an informative prior because apparently the researcher already knows a lot. (If such knowledge is available, it would actually be very strange to ignore it, which is what is basically done in the frequentist approach!) An informative prior will have a rather strong influence on the posterior, especially if there are not too many data. If, however, one knows very little about θ, one can give a so-called uninformative, or flat prior, for instance a uniform distribution where a parameter can take on every value in an interval specified by the researcher. In most cases, a so-called weakly informative prior is the best (McElreath, 2016). It means that one should incorporate the little knowledge that one may have, for instance that a parameter is non negative, or has an upper bound, in the prior. This prevents overfitting (meaning that a model is well capable to explain the available data but not to predict new data). Incidentally, an informative prior has a substantial effect on the posterior only if there is not much information in the data. With more data available the likelihood becomes dominant quickly. The upper panel of Fig. 3 gives an impression of the effect of the prior standard deviation on the posterior, given a certain likelihood function, while the lower panel of Fig. 3 shows the effect of number of data points given an informative prior.

As mentioned above, for kinetic problems a normal (Gaussian) distribution can be assumed for parameters such as rate constants and activation energy. As McElreath (2016) puts it: "A Gaussian distribution is the most natural expression of our state of ignorance, the least surprising and least informative assumption to be made". The standard deviation (see Fig. 1) is one of the parameters to be estimated from uncertain data and therefore the standard deviation that characterizes uncertainty in the data is itself uncertain! A standard deviation cannot be negative, so the prior should be bounded at zero. A distribution that allows for a broad range of possible values while becoming less likely at extremes suits the best. Experience shows that a half-Cauchy distribution for the standard deviation σ functions well as a weakly informative, regulating prior (Lambert, 2018; McElreath, 2016); the half-Cauchy distribution resembles a t-distribution with thick tails bounded at 0 (see Fig. 2D); an alternative distribution could be the exponential distribution which is also bounded at 0 but does not have thick tails.

It is perhaps interesting to note that the frequentist method can actually be considered as a special case of the Bayesian approach, namely with a completely uninformative flat prior, so that the posterior is only determined by the likelihood (i.e., the data), hence with only the likelihood function as information (but remember that the likelihood function as a function of parameters is not a valid probability distribution). One then searches for the parameter value that makes the

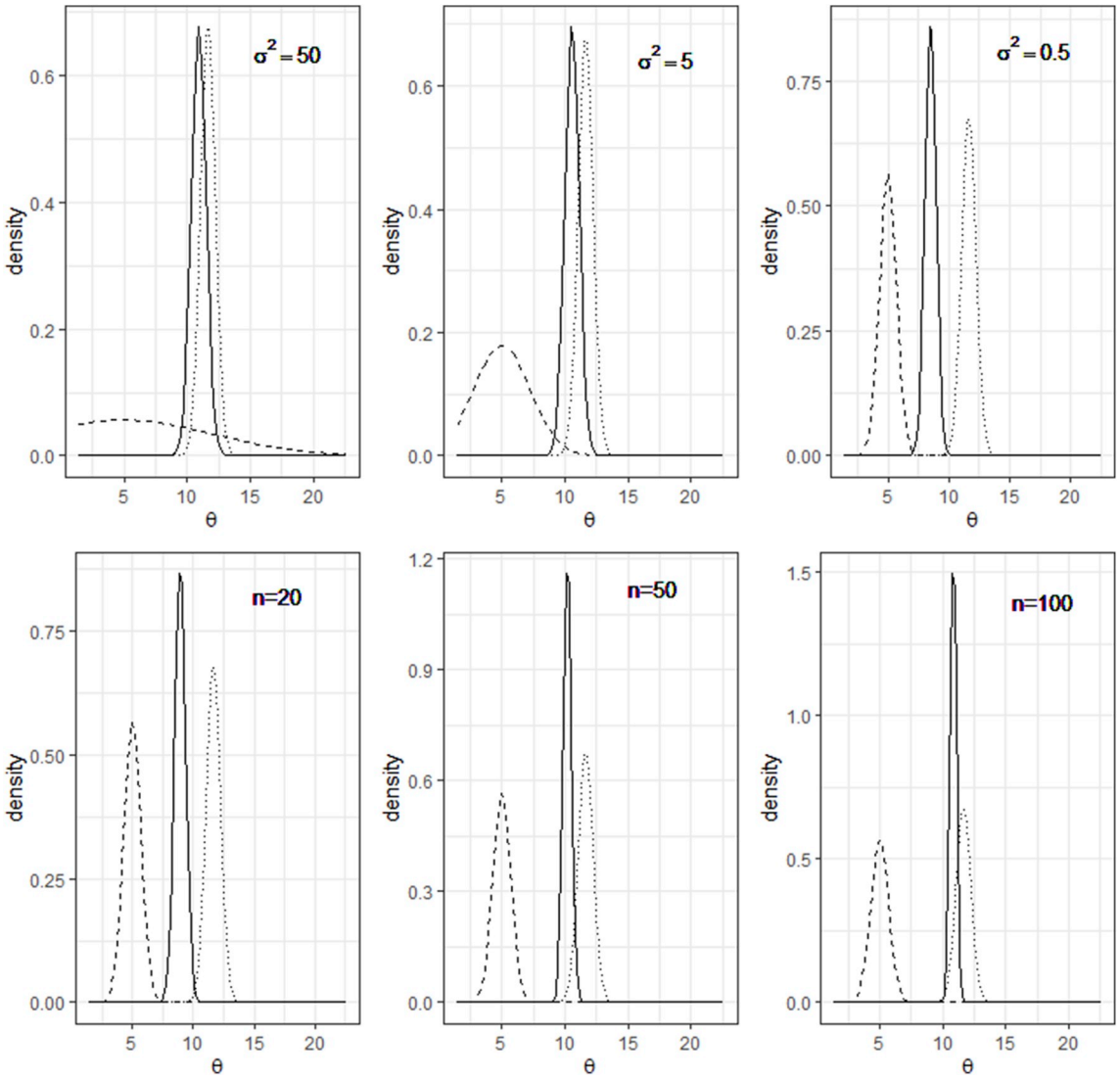

Fig. 3. Upper panel: effect of a prior normal distribution ($\mu = 5$ and varying σ^2_{prior} (dashed line) on the posterior outcome (solid line) with a normal likelihood distribution ($n = 20$ datapoints with $\mu = 12$ and $\sigma^2 = 7$ (dotted line). Lower panel: effect of n datapoints (normal likelihood with $\mu = 12$ and $\sigma^2 = 7$) (dotted line) with a prior $\mu = 5$ and $\sigma^2 = 0.5$ (dashed line) on the posterior density (solid line).

obtained data most likely; however, it does not result in a posterior distribution if no prior is specified. A prior, however uninformative, needs to be specified in the Bayesian approach. If prior information about parameters is available but ignored, it would be appropriate to mention why it is ignored, which makes the omission of the prior actually an argument against the frequentist approach. In the Bayesian approach it cannot be ignored while the researcher is also forced to think about the choice for a likelihood function that describes the data.

To be complete, the frequentist process of searching for the most likely parameter without specifying a prior is called maximum likelihood estimation (MLE) and is also the basis for the commonly applied ordinary least squares (OLS), *if several assumptions are fulfilled* (Van Boekel, 1996). This can be illustrated as follows. If the likelihood of one datum point y is $f(y|\theta)$, then the likelihood of all datapoints y_i is found from a multiplication (symbol $\prod_{i=1}^{n}$) of the likelihood of each individual data point (this follows from probability theory):

$$l\left(y_i|\theta\right) = \prod_{i=1}^{n} f\left(y_i|\theta\right) \tag{11}$$

The value of θ that maximizes this likelihood function makes the data as they have been observed most likely. In the search for this value it is mathematically more easily done by taking logarithms, because minimizing the log-likelihood ($L(y_i|\theta) = \ln(l(y_i|\theta))$ gives the same result as maximizing the likelihood:

$$L\left(y_i\middle|\theta\right) = \ln\left(l\left(y_i\middle|\theta\right)\right) = \sum_{i=1}^{n} \ln\left(f\left(y_i|\theta\right)\right) \tag{12}$$

If we take the normal distribution for $f(y_i|\theta)$ (see equation (5)), it follows that:

$$L(y_i|\theta) = \ln(l(y_i|\theta)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)\right) \tag{13}$$

The last part in the right hand side of this equation contains the sums of squares between data y_i and model μ_i and minimizing this sums-of-squares is equivalent to minimizing the ln(likelihood), which in turn is equivalent to maximizing the likelihood. So, if the normal distribution describes the variation in the data with a constant standard deviation, then it follows that minimization of the sums-of-squares is equivalent to MLE, and this is then the rationale for using OLS in the frequentist approach. If OLS is to be interpreted in the Bayesian

framework, it could be seen as using a uniform prior from to + ∞ for the regression coefficients and a uniform prior from 0 to + ∞ for the standard deviation. MLE is the result of *maximizing* over all possible parameter values, whereas the Bayesian way results in *integrating* over all possible parameter values. Inference following MLE results, therefore, only in point estimates, while the Bayesian approach results in a probability distribution with much more information.

All in all, there is a different perspective on probability in the Bayesian and frequentist approach; Bayesians consider probability as a measure for the degree of belief, frequentists make probability statements about repeated sampling from a population.

So, now we have the basis for doing Bayesian regression. These are the four basic steps in Bayesian data analysis (after Muth et al. (2018)).

1. specify the model as a deterministic equation, whereas the likelihood and the prior (one for each parameter) are specified as a statistical model
2. Estimate the model parameters (including σ) and check for convergence
3. Check the fit of the model
4. Interpret the results

Incidentally, these steps are also appropriate in the frequentist framework, except for the prior and likelihood specification. A detailed checklist for these steps in the Bayesian case can be found in Depaoli and Van de Schoot (2017).

2.3. Software for Bayesian estimation

Before we show an example, a few words on the software that is used nowadays in Bayesian estimation: what does it actually do? For all but very simple cases, the formula in Bayes theorem, equation (9), when coupled to statistical models, is not analytically solvable anymore, it will involve enormously complicated integrals appearing in the denominator. So, to find this posterior distribution, the best way is to approximate it via numerical techniques. A relatively simple method is to use a quadratic (or Gaussian) approximation to the posterior (Looking at equation (13), the part $(y-\mu)^2$ causes the quadratic shape, and taking the exponential of that leads to the typical bell shape of a normal distribution.). This approximation is strictly speaking only allowed when the posterior has a multivariate normal distribution (which in many cases may not be a bad assumption). The software then searches for the mode of the distribution as specified by the prior and the likelihood, and then it searches for the curvature around this mode. For instance, the R function "map", (map = maximum a posteriori) as part of the package "rethinking" can be used to this goal (McElreath, 2016). A more rigorous approach (because no assumptions need to be made about the posterior) is via Monte Carlo simulations and variants thereof. What the software does is to explore the region of the posterior (as specified by the prior and the likelihood) via random walks and drawing many samples from the target posterior distribution; the result is that these samples in the end represent the unknown distribution. Techniques that are used in this respect are Gibbs sampling, Metropolis Hasting sampling, Hamiltonian Monte Carlo and the No_U Turn sampler (NUTS). These are commonly summarized with the term Markov Chain Monte Carlo methods (MCMC). A Markov chain is a sequence of states in which the location of the next state depends on the current state of the chain. Currently, the most powerful and efficient algorithm is the Hamiltonian Monte Carlo algorithm, and its extension NUTS, which is used in Stan (Bürkner, 2018a; Gelman, Lee, & Guo, 2015; Kruschke, 2015; Lambert, 2018; McElreath, 2016). A very instructive tutorial explaining MCMC can be found on the internet: http://www.flutterbys.com.au/stats/tut/tut4.3.html. We will apply Stan in the kinetics example later on. Those who are interested in background reading about MCMC are referred to Betancourt (2017).

In order to use the MCMC concept fruitfully in practice, some diagnostics are needed to check that the MCMC procedure has converged. Whether or not the number of iterations in MCMC is sufficient can be checked by looking at the so-called trace plots: they should show no trends and should stabilize around a baseline with some random noise. MCMC simulates a distribution in proportion to the density but a problem can be that the consecutive samples are very similar. The solution for this is to skip some samples during the iterations, this is called "thinning the sample set", so that the remaining samples will be more independent of each other. So, the modeller can indicate a thinning factor (for instance, a thinning factor of 10 means: retaining every 10th sample; as a consequence more iterations are needed in order to obtain sufficient samples). Then, starting values for drawing samples are needed (either provided by the modeller, or guessed by the software) and it may take some time before the software ends up in the right region. Therefore, a burn-in or warm-up interval is defined, the samples of which will be discarded, a rule of thumb is that at least the first 10% of the samples should be ignored, but usually 50% is used to be sure. MCMC algorithms do not have a stopping criterion, so the modeller should indicate the maximum number of iterations (trace plots will indicate whether or not the number of iterations was sufficient, if not, this number should be increased). In order to speed up the sampling process, more chains are used (e.g., 3 or 4) at the same time. Then, there is the Gelman and Rubin diagnostic called R-hat ($\hat{R}$), which measures the ratio of the variance among chains and that within chains. This ratio should become very close to 1, otherwise the chains are not mixed well and did not converge to a stationary value; it basically measures whether or not chains have covered the whole space of the posterior. Another diagnostic is the "number of effective sample size". This has to do with autocorrelation of the sampling process. It compares the number of independent draws with the same estimation accuracy of correlated draws. The sampling algorithm rejects some samples according to a certain criterion and calculates how many remaining samples are effectively used. Therefore, the number of effective samples is usually less than the number of total samples. There is no absolute criterion for the effective sample size, if it is less than 10% of the total sample size ones should be careful and change some settings. In any case, it is essential to do some MCMC diagnostics before interpreting MCMC results. The example below will demonstrate their use.

To some, the MCMC approach may seem as a black box approach, but it is not. It is very well described and documented, and accepted as a valid method, and supported by renowned statisticians. It is important to realize that priors are needed for MCMC sampling, even if no prior information is available (i.e., a noninformative prior should always be given as the least) and that distinguishes it from maximum likelihood estimation. Dedicated software programmes are available for MCMC sampling, like WINBUGS, JAGS, Stan and several algorithms in R, most of them come at no cost. An internet search on "software for MCMC" will give the interested reader an overview.

3. A kinetics example

For the remainder of the article, the above discussed concept is applied; a very simple kinetics example was chosen so that the focus lies on the Bayesian aspect rather than on the kinetic problem itself, in other words the kinetic problem is considered trivial. In due course we will focus much more on how the Bayesian approach can help to unravel complicated kinetic problems. References for the R version, R Studio and packages used are given in the supplemental material. Instructions on how to install R, RStudio and Stan, as well as several other R packages, can be found on internet. The actual R codes for the calculations can be found on the author's GitHub repository: https://github.com/TinyvanBoekel/Trends.

The example chosen is about the formation of furan in soy sauce at 70 °C for which a linear relation with time was found, i.e., a zero-order reaction as in equation (2). Furan is formed in the Maillard reaction (Huang & Barringer, 2016) as one of the reaction paths. Much can be

said about the kinetics of this reaction (e.g., see Chapter 8 in Van Boekel (2008)) but the main purpose here is to show how we can analyse such data in the Bayesian way. The example was chosen as a typical food science problem with not too many data points; information about experimental variation could not be retrieved from the article. Such a dataset was chosen on purpose rather than an extended dataset with replications. Unlike the frequentist framework, Bayesian statistics is not based on large samples, so a small sample size without replicates was selected as a showcase. As a kinetic problem it is trivial, it is chosen for simplicity to be able to focus attention on the Bayesian aspects of regression. As mentioned, a zero-order model is proposed:

$$c_i = c_0 + k_r \cdot t_i \tag{14}$$

The outcome of the classical (frequentist) approach via OLS is shown, for reference, in Table S1 (provided in the supplemental material). This table shows, among other things, the point estimates of the slope and intercept plus 95% confidence intervals; it also shows the p-value, which should be read as the probability that the null-hypothesis is true. As a reminder: in classical (frequentist) statistics the null hypothesis is always that there is no effect. In this case the null hypothesis is that there is no effect of the predictor time on furan concentration (i.e., that the slope is zero) and that the intercept equals zero. The probability shown for the slope to be zero is low ($p < 0.001$), so it is highly unlikely that the null hypothesis of no effect of heating time on furan formation is true; on the other hand, the probability that the intercept is zero is quite high ($p = 0.59$). This is how p-values should be interpreted: as statements about the null hypothesis. With respect to the 95% confidence intervals shown in Table S1, the interpretation is that if we would repeat the analysis many times, the parameter would be in the interval shown for 95% of the repeated analyses. It is not a statement about the parameters.

Knowing the conventional OLS outcome, a Bayesian analysis follows next, using the above mentioned 4 steps.

3.1. Propose models

As with the frequentist approach, the two parameter deterministic model shown in equation (14) is used, to estimate the initial concentration c_0 and the reaction rate constant k_r. This deterministic model needs to be coupled to a statistical model. As discussed above, for chemical and physical measurements, a normal distribution is to be expected for the data (furan measurements), so that will be the likelihood function. The data are coupled to the expected value μ_i, which is in turn predicted by the deterministic model. Also for the two parameters c_0 and k_r, a normal distribution is proposed as priors, each characterized by a mean and a standard deviation, while the (assumed constant) standard deviation is supposed to be distributed as a half-cauchy distribution. Having proposed the deterministic and statistical relationships, we have to provide numerical values to the prior distributions before we can proceed with the actual regression (not for μ_i because it will be estimated from the parameters). We will try weakly informative priors to start with, i.e., priors that hold some useful information without strongly influencing the final parameter estimates (Depaoli & Van de Schoot, 2017). We will come back to the effect of prior choice later on.

$$\begin{aligned} &c_i \sim \mathcal{N}(\mu_i, \sigma) \\ &\mu_i = c_0 + k_r \cdot t_i \\ &c_0 \sim \mathcal{N}(0{,}100) \\ &k_r \sim \mathcal{N}(1{,}10) \\ &\sigma \sim \text{half} - \text{cauchy}(25) \end{aligned} \tag{15}$$

3.2. Estimate the model parameters

The R package "brms" (Bayesian Regression modeling using Stan) was used for Bayesian estimation, v. 2.8.0. brms acts as intermediate between R and Stan (Bürkner, 2018). Stan does the actual MCMC calculations (Gelman et al., 2015). The following code shows how simple the programming actually is; what needs to be specified are the data ("furandata""), the likelihood ("family = gaussian" with the deterministic model expressed as: "formula = furan ~ 1 + time", priors for the "intercept", slope ("b") and standard deviation ("sigma"). What also needs to be specified is the number of chains, number of iterations and how many warm-up samples need to be discarded; the adapt_delta setting helps to converge. Here is the brms code.

```
library(brms)


time <- c(0,30,60,90,120,150) # in min
furan<- c(0,54,84,147,177,225) # in microgram/L
furandata <- data.frame(time, furan)


furanfit <-
  brm(data = furandata, family = gaussian,
      formula = furan ~ 1 + time,
      prior = c(set_prior("normal(0, 100)", class = "Intercept"),
                set_prior("normal(1, 10)", class = "b"),
                set_prior("cauchy(0,25)", class = "sigma")),
      chains = 4, iter = 4000, warmup = 2000,
      control = list(adapt_delta =0.95),
      inits = 0, seed=15, file="furanfit")
furanpost <- posterior_samples(furanfit)
```

Running this code within brms in R, the package will pass on the data, the priors and the model to Stan, and Stan will return the result of the MCMC calculations to R for further analysis.

3.3. Check the model fit

The first thing to do after MCMC estimation is to check whether the chains have converged, see the trace plots shown in the upper panel of Fig. S1 (supplemental material). These have the typical 'caterpillar" shape, arising from the mixing of the chains. It appears that the chains are well mixed and have converged to stable values. The lower panel of Fig. S1 (supplemental material) shows the histograms for a first impression of the posterior distribution of the parameters, while the scatter plots give an impression of the correlation between parameters, which are, in addition, shown numerically as correlation coefficients in the same Figure.

A numerical impression of the results is shown in Table S2 (supplemental material). Two parameters in this Table tell something about the convergence, the n_{eff} which should be > 100 and $\hat{R}$ which should be close to 1. Next to that, the Table also shows the numerical estimates of the intercept and slope plus the uncertainties involved. Note that these numbers should be interpreted as summaries of the posterior densities, representing central tendency measures for the posterior rather than point estimates (Depaoli & Van de Schoot, 2017).

Though the numerical summaries from the Bayesian regression are not exactly the same as the point estimates from the frequentist analysis (compare Tables S1 and S2 in the supplemental material), they are quite close. However, the difference with least-squares regression is that we have obtained a posterior distribution with a lot of information, whereas least-squares regression only gives a point estimate. Table S3 (supplemental material) shows the first six entries of what is inside the posterior (out of the 2000 entries, remember that 4000 samples were

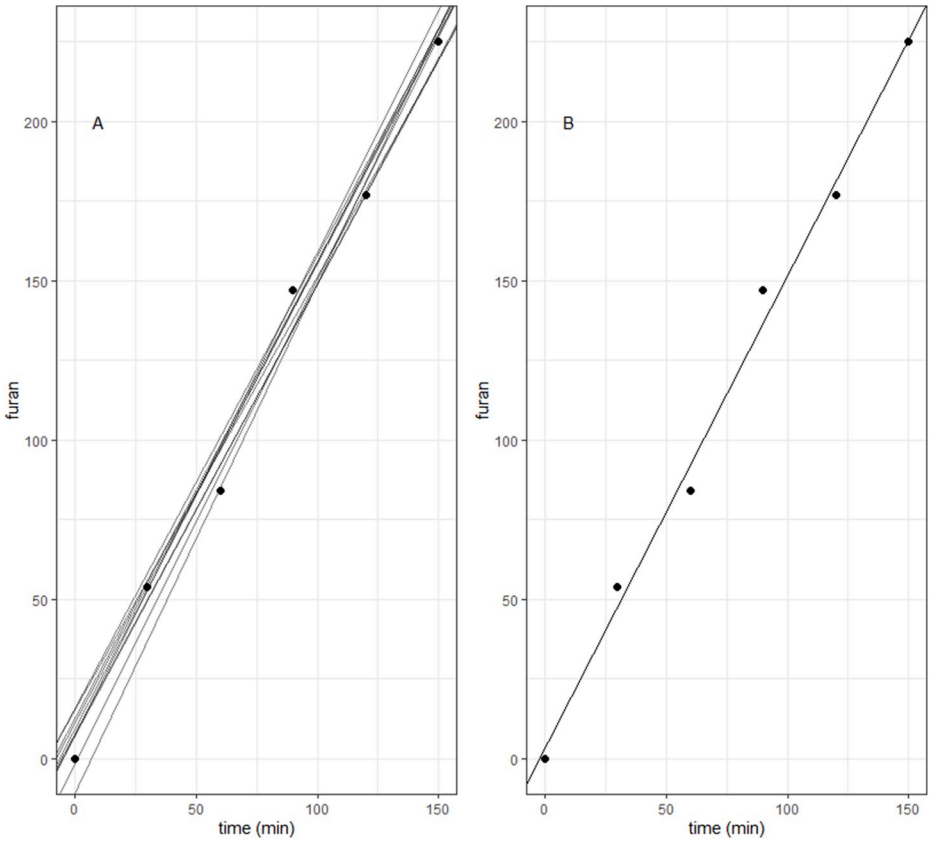

Fig. 4. Plot showing regression lines for the first 10 rows of the posterior (A) and the regression line obtained by applying the mean of all 2000 intercepts and slopes (B).

obtained with the first 2000 warmup samples discarded). Each line in the table shows one particular result of the 2000 MCMC estimations. With this information in the posterior, regression lines could be calculated, but also covariances and correlations between parameters (as shown in the lower panel of Fig. S1 of the supplemental material).

Fig. 4A gives a first impression of the variation in parameter estimates by plotting 10 regression lines from 10 entries in the posterior, while Fig. 4B shows the line representing the mean of all the 2000 posterior samples. Note that this line is NOT the result from minimization of sums-of-squares! Rather, it represents the most likely combination of the parameters that are contained within the posterior. Also note that this line is not a prediction of the model but a fit to the experimental data; McElreath (2016) calls this retrodiction, which seems an appropriate term as it shows how the model complies with the data in retrospect.

3.4. Interpret the results

Having the posterior at our disposal, and knowing that the estimation procedure went apparently well, we can do all kinds of analysis with it. One is to show credibility and prediction intervals. (To distinguish the Bayesian approach from the frequentist approach, a Bayesian confidence interval is called a credibility interval.) Credibility intervals show how confident we can be in estimating the mean μ. Let's do that first for a certain time point, for instance, what is the expected value of furan at 100 min? The equation that we can derive from the posterior for the regression is $\mu_{100} = 3.06 + 1.48 \times 100 = 151.35$ μg/L. But since there is variability in c_0 and k_r (as quantified in the posterior) there must also be variability in μ_{100}, and this variability can be calculated right away by sampling from the posterior: see Fig. 5A. We can also establish a Highest Posterior Density Interval (HDPI) at say, 95% credibility (could be any number, to be chosen by the researcher); this information is supplied also in Table S2 (supplemental material), and could be plotted if so desired.

The density profile reflected by the solid line in Fig. 5A thus shows the range where we may expect the mean μ_{100} to be for $t = 100$ min; if so desired, we could also indicate the range for a certain probability (e.g., 50%, or 95%, or whatever value one finds interesting). However, this range applies to the **mean** calculated at t = 100 min, not to a **future prediction** at $t = 100$ min, which is obviously more interesting if our final goal is to predict. So, what we need to produce next is a prediction interval that expresses not only how confident we can be in estimating the mean, but also in how confident we can be in predicting new data at $t = 100$ min. To do that, we need to take into account the variation present due to data collection; this latter variation is characterized by the parameter σ, for which we also have estimates as shown in Table S3 of the supplement. The result of that calculation is represented as the distribution indicated by the dashed line in Fig. 5A. In summary, for prediction purposes we need to take into account two sources of variation, one being the uncertainty in estimating the mean (due to uncertainty in the parameters) and on top of that the variability due to experimental error (characterized by sigma). Because there is more variation involved in prediction this is obviously a broader distribution but with the same mean, as shown in Fig. 5. Generally, instead of looking at a specific value like μ_{100}, we can do these calculations for the whole spectrum covered by the regression line. This is shown in Fig. 5B.

Next to the prediction line, Fig. 5B also shows that the measured

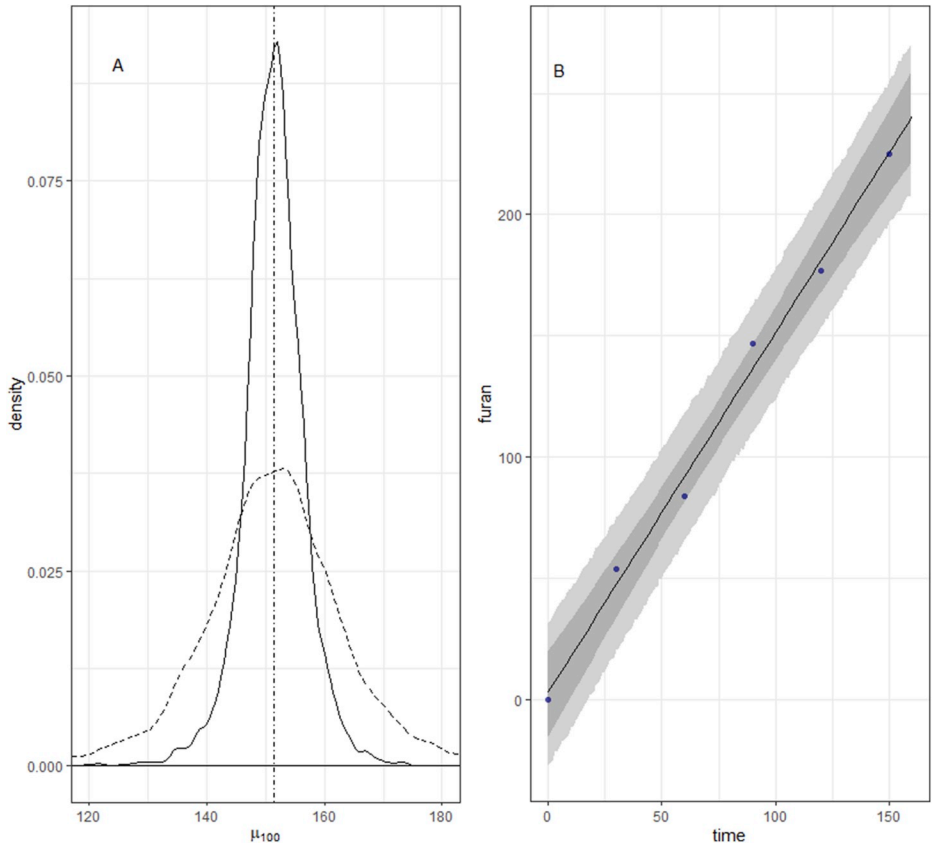

Fig. 5. A: Density of μ_{100} for the expected furan value at $t = 100$ min (solid line) and for a new predicted value (dashed line). The vertical dashed line indicates the mean value of μ_{100} (A). B: Regression line for the formation of furan, with the 95% credibility limits for the mean (dark grey area) as well as 95% prediction limits for new values (light grey area).

values fall within the prediction range. This is one of the so-called posterior predictive checks (PPC): if the measured values would fall outside that range, it would raise suspicion about the predictive performance of the model. It would of course be better to have an independent new sample to compare with the model prediction but that is not always feasible. There is another option to explore the results of the MCMC analysis in the posterior, including posterior predictive checks, via the package "shinystan", as shown for instance by Muth et al. (2018). It gives an interactive graphical interface to visualize, summarize and diagnose the MCMC analysis. One can export tables and graphics with this package. It is not shown here but the code to launch this package is given with the rest of the code on the author's Gihub repository. There are actually many more options to explore the posterior. In this respect, a very instructive tutorial for linear regression in the Bayesian way can be found on internet: http://www.flutterbys.com.au/stats/tut/tut7.2b.html.

It may be instructive to compare the obtained marginal posteriors for the parameters with the priors that were chosen before the analysis to see how much we have learned from the data (Fig. 6).

The posterior densities (which were displayed as histograms in Fig. 1 in the supplemental material) are much tighter than the priors. The prior for k_r can hardly be seen on the scale of the posterior. It goes to show that much information has been gained from the data. It appears that the posteriors for c_0 and k_r are approximately normally distributed, while the posterior for σ is skewed to the right. This is usually so for the standard deviation because it is bounded by zero as the lower limit (McElreath, 2016). The posterior distribution for c_0 shows that zero is definitely within the interval and this is as expected because at time zero furan is not yet formed; however, the posterior distribution also shows that there is considerable uncertainty in this parameter, in other words it cannot be very well estimated from the data available. In principle, of course, an initial concentration cannot be negative, so the posterior should not be interpreted that negative values are possible for this parameter, it just indicates that the data do not allow it to be estimated as being exactly 0. The posterior for k_r shows clearly that it does not contain zero in its interval, and so there is a definite effect of heating time on furan formation, but again with some variation around the estimated value. A rate constant cannot be negative from a physical point of view but it can carry a negative sign if there is a degradation reaction rather than a formation reaction. In general, it is advised to let the data decide whether the slope should be positive or negative. In that sense, a normal distribution, which runs in principle from $-\infty$ to $+\infty$, as a prior for a rate constant makes sense. If a parameter really needs to be restrained to positive values, one can use a distribution that only covers positive values, such as the gamma distribution, the cauchy distribution, the exponential distribution or the log-normal distribution.

3.5. Dealing with outliers

Although there is no strong indication for outliers in the data set (see Fig. 4B), it might be interesting to check whether or not a t-distribution rather than a normal distribution would give different results. The t-distribution (also called the student distribution) is much more robust to outliers than the normal distribution (Kruschke, 2015). The t-distribution has one parameter more, namely the parameter ν, in the frequentist framework known as degrees of freedom. In fact, the normal distribution is a special case of the t-distribution with $\nu = \infty$. It is very easy in brms to do regression with the t-distribution; if so desired we can give a prior for the parameter ν or accept the default prior which is gamma(2,0.1). The following code shows how easy it is to move from a

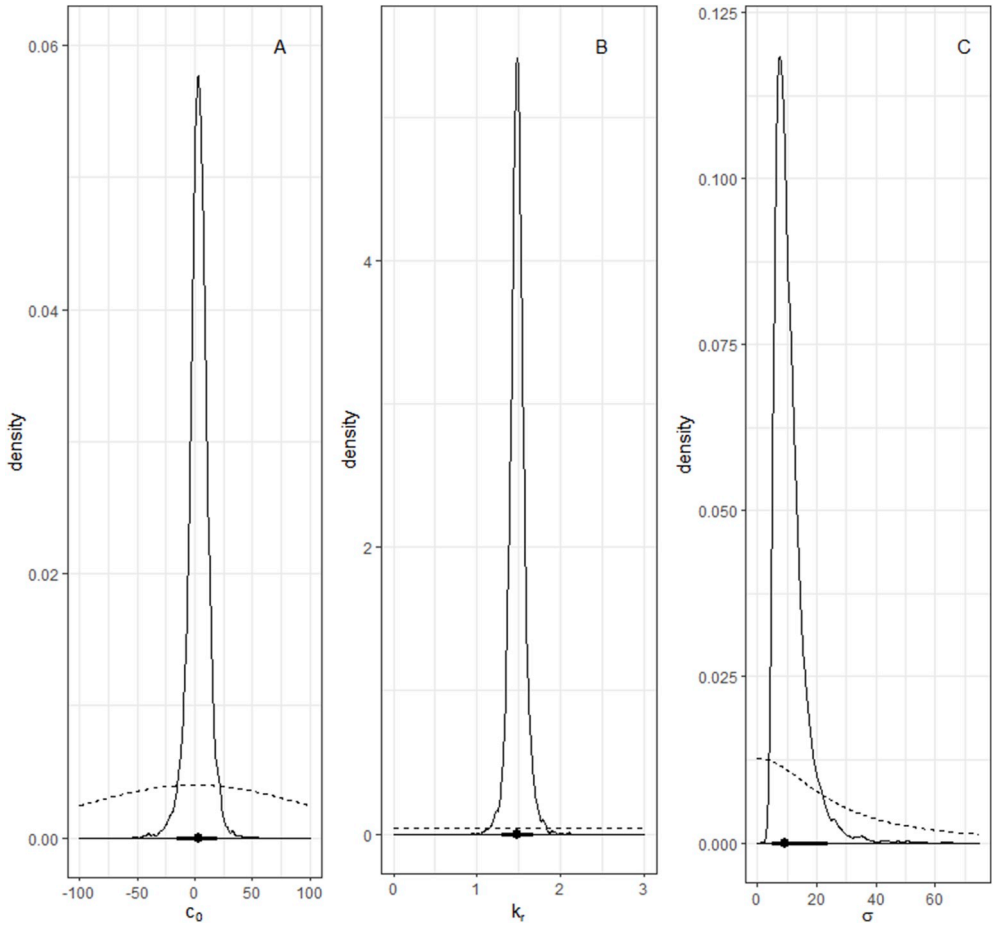

Fig. 6. Density plots of c_0 (A), k_r (B) and σ (C). Solid lines: posterior, dashed lines: prior. The bold line indicates the 95% HDI (highest density interval) with the mean as big dot.

normal distribution to a student distribution; note that an extra prior is given for the ν parameter.

```
furanfit_student <-
  brm(data = furandata, family = student,
      formula = furan ~ 1 + time,
      prior = c(set_prior("normal(0, 100)", class = "Intercept"),
                set_prior("normal(1, 10)", class = "b"),
                set_prior("gamma(2,0.1)", class = "nu"),
                set_prior("cauchy(0,25)", class = "sigma")),
      chains = 4, iter = 4000, warmup = 2000,
      control = list(adapt_delta =0.95),
      seed=15, file="furanfit_student")
```

The MCMC procedure went well as shown by the trace plots in Fig. S2 of the supplement. The numerical summaries are shown in Table S4 for comparison with Table S2 (both shown in the supplement). The estimates are slightly different but not drastically, the estimate for the ν parameter is 21 with quite some variation, indicating that the assumed distribution is not completely normal but not disturbingly so. According to Kruschke (2015), a ν value $>$ 30 gives virtually a normal distribution.

3.6. Choice of priors

Finally, it may be instructive to investigate how strong the effect of priors can be on the final outcome, i.e., the posterior; this is called a prior sensitivity analysis (Depaoli & Van de Schoot, 2017; Korner-Nievergelt et al., 2016). Since the number of data points (six) in the present furan example is not very high, there could be some effect of the choice of prior on the final outcome (compare Fig. 3). However, there appeared to be an effect only when very informative priors were used for the parameters, i.e. when the standard deviations for the priors were chosen to be $<$ 0.5 for the prior on k_r and $<$ 50 for the prior on c_0. The resulting fits were completely out of range in those cases, so it was immediately obvious that underfitting occurred: the strong priors prevented the model from learning from the data (McElreath, 2016). By increasing the standard deviation, a point is reached where the prior has no effect anymore on the numerical summaries. The choice made above for the priors in the case study can thus be characterized as weakly informative. One may wonder why not always use non-informative priors? The reason is that priors that are slightly informative prevent overfitting (in the sense that models get too excited about the data, as McElreath (2016) puts it). Weakly informative priors also have a stabilizing effect on the MCMC simulation (Korner-Nievergelt et al., 2016).

Some guidelines for choosing priors are (see also Gelman, Simpson, and Betancourt (2017); Depaoli and Van de Schoot (2017)):

1. Set priors according to what you think is real but leave room for some doubt and give values in the same order of magnitude as what you are trying to predict. Don't use flat, uninformative priors and beware of uniform priors (they are usually unreal) unless you have a good reason to. Most importantly, a prior must make sense to you, it should not just be an arbitrary choice. Make a plot of the chosen priors to visually check whether or not they display what you think is reasonable.

2. Be explicit about why you choose a certain prior so that it is debatable for your peers
3. Do a prior sensitivity analysis to get a feel for their influence.
4. Do not change a prior just to obtain a result that seems desirable, that would come close to data manipulation. If it turns out that the choice of a prior has a strong influence on the results, this should be discussed in detail.

What is not shown here is that there is no principal distinction between regression of linear and nonlinear models in the Bayesian approach (McElreath, 2016). With OLS, nonlinear regression needs a different approach than linear regression (Van Boekel, 2008). However, depending on their complexity, nonlinear models may sometimes cause numerical difficulties. The brms package can handle nonlinear models (Bürkner, 2018b).

4. Conclusion: pros and cons of the Bayesian approach

In conclusion, it is hoped that this paper has contributed in gaining more insight in the Bayesian approach of regression of kinetic models, and the advantages it may have. In summary the main advantages of Bayesian regression, in view of the author, are:

1. Strong emphasis on uncertainty and how we gain knowledge in science (i.e., decrease uncertainty) by learning from data and models, in particular by visualizing them in density plots and credibility and prediction intervals. It makes parameters better interpretable. Another thing, not shown here, is that Bayesian modeling makes parameter identifiability easier, or rather that it is easier to detect with Bayesian modeling when parameters are not identifiable (Hines, Middendorf, & Aldrich, 2014).
2. Straightforward interpretation of credibility (confidence) and prediction intervals: probabilities express the uncertainty in a quantitative and natural way.
3. Correct and immediate propagation of parameter uncertainties when they are used in predictions, all the necessary information for that is in the posterior; no intricate and hard to explain formulas are needed as with OLS regression (De Levie, 2012).
4. Any proper deterministic model connected to the proper probability distribution can be investigated via MCMC, linear or nonlinear, making it a far more general approach than OLS (Hobbs & Hooten, 2015). So, we could do basically the same analysis as shown here for a linear model for a nonlinear first-order model, for instance, or any other model that the researcher deems appropriate. Of course, it remains the responsibility of the researcher to check that the assumptions made make sense.
5. Although not shown in this paper, Bayesian methods are easily extended to more complex models. Also, other likelihoods than the normal distribution are easily implemented, the MCMC approach can handle that just as easily.
6. Also not shown in this paper but worthwhile to mention is that Bayesian methods work very well with multilevel models (sometimes also called mixed effect level models), i.e., models that share parameters on various levels (e.g., at population and individual levels); see, for instance, McElreath (2016), Lambert (2018). A recent example using brms is given by Nalborczyk, Batailler, Loevenbruck, Vilain, and Bürkner (2019), be it on the very different research discipline linguistics (the levels concern gender effects on vowel production in standard Indonesian). A case more related to food science could be with models for growth or inactivation of microorganisms, to separate population effects from individual variation caused by, for instance, different days of inoculation.

As always, there are also disadvantages:

1. One has to do some programming in a language such as R
2. One has to choose priors and likelihoods. While that may be a challenge at first, the disadvantage turns into an advantage when the modeller is forced to think hard about the model and the underlying assumptions.
3. One has to have some idea of probability theory.
4. Computations take time, MCMC sampling from the posterior requires a substantial number of iterations (the rather simple example here takes already several minutes of calculation). This "problem" will diminish with ever increasing computer speed.

In the author's view, the advantages outweigh the disadvantages considerably. Even though the point estimates of Bayesian and frequentist regression do not differ substantially, at least not in the example used, there may be the advantage of coming closer to the way people grasp statistics intuitively. The Bayesian approach requires a different way of thinking and a different mindset that may help in this respect. Looking at the rapid increase in publications about this topic, it is inevitable that this approach will also enter chemical kinetics and food science literature soon, so we better be prepared.

Author note

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors. Declarations of interest: none.

Acknowledgements

The author acknowledges gratefully comments from dr. Matthijs Dekker and Prof. dr. Karin Schroen on an early version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tifs.2020.02.027.

References

Baker, M. (2016). Statisticians issue warning over misuse of P values. *Nature, 531*(7593), 151. https://doi.org/10.1038/nature.2016.19503.

Baldwin, S. A., & Larson, M. J. (2017). An introduction to using Bayesian linear regression with clinical data. *Behaviour Research and Therapy, 98*, 58–75. https://doi.org/10.1016/j.brat.2016.12.016.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. Retrieved from http://arxiv.org/abs/1701.02434.

Box, G. E. P., & Draper, N. R. (1965). The Bayesian estimation of common parameters from several responses. *Biometrika, 52*(3–4), 355–365. https://doi.org/10.1093/biomet/52.3-4.355.

Bryan, J. (2018). Excuse me, do you have a moment to talk about version control? *The American Statistician, 72*(1), 20–27. https://doi.org/10.1080/00031305.2017.1399928.

Bürkner, P.-C. (2018a). Advanced Bayesian multilevel modeling with the R package brms. R J. 10(1), 395-411, Retrieved from http://arxiv.org/abs/1705.11123.

Bürkner, P.-C. (2018b). Estimating multivariate models with brms. Retrieved from https://cran.r-project.org/web/packages/brms/vignettes/brms_multivariate.html.

Carreno, J. J., Lomaestro, B., Tietjan, J., & Lodise, T. P. (2017). Pilot study of a Bayesian approach to estimate vancomycin exposure in obese patients with limited pharmacokinetic sampling. *Antimicrobial Agents and Chemotherapy, 61*(5), https://doi.org/10.1128/AAC.02478-16 e02478–16.

De Levie, R. (2012). *Advanced Excel for scientific data analysis* (3rd ed.). New York: Oxford University Press p. 646.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240–261. https://doi.org/10.1037/met0000065.

Eguchi, T. (2008). An introduction to Bayesian statistics without using equations. *Marine Turtle Newsletter,* (122), 1–5.

Galagali, N., & Marzouk, Y. M. (2015). Bayesian inference of chemical kinetic models from proposed reactions. *Chemical Engineering Science, 123*, 170–190. https://doi.org/10.1016/j.ces.2014.10.030.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530–543. https://doi.org/10.3102/1076998615606113.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood

in the context of the likelihood. *Entropy, 19*(10), https://doi.org/10.3390/e19100555.
Goula, A. M., Prokopiou, P., & Stoforos, N. G. (2018). Thermal degradation kinetics of L-carnitine. *Journal of Food Engineering, 231*, 91–100. https://doi.org/10.1016/j.jfoodeng.2018.03.011.
Grainger, M. N. C., Owens, A., Manley-Harris, M., Lane, J. R., & Field, R. J. (2017). Kinetics of conversion of dihydroxyacetone to methylglyoxal in New Zealand mānuka honey: Part IV - formation of HMF. *Food Chemistry, 232*, 648–655. https://doi.org/10.1016/j.foodchem.2017.04.066.
Halpern, A. M., Frye, S. L., & Marzzacco, C. J. (2018). Scientific data analysis toolkit: A versatile add-in to microsoft excel for windows. *Journal of Chemical Education, 95*(6), 1063–1068. https://doi.org/10.1021/acs.jchemed.8b00084.
Hines, K. E., Middendorf, T. R., & Aldrich, R. W. (2014). Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *The Journal of General Physiology, 143*(3), 401–416. https://doi.org/10.1085/jgp.201311116.
Hites, R. A. (2017). Calculating the confidence and prediction limits of a rate constant at a given temperature from an Arrhenius equation using excel. *Journal of Chemical Education, 94*(3), 398–400. https://doi.org/10.1021/acs.jchemed.6b00842.
Hobbs, N., & Hooten, M. (2015). *Bayesian models: A statistical primer for ecologists.* Princeton University Press pp. 1–299.
Huang, X., & Barringer, S. A. (2016). Kinetics of furan formation during pasteurization of soy sauce. *Lebensmittel-Wissenschaft und -Technologie- Food Science and Technology, 67*, 200–205. https://doi.org/10.1016/j.lwt.2015.11.053.
Korner-Nievergelt, F., Roth, T., Von Felten, S., Guelat, J., Almasi, B., & Korner-Nievergelt, P. (2016). *Bayesian data analysis in ecology using linear models with R, BUGS and Stan.* Elsevier.
Krauss, M., Tappe, K., Schuppert, A., Kuepfer, L., & Goerlitz, L. (2015). Bayesian population physiologically-based pharmacokinetic (PBPK) approach for a physiologically realistic characterization of interindividual variability in clinically relevant populations. *PloS One, 10*(10), e0139423. https://doi.org/10.1371/journal.pone.0139423.
Kruschke, J. (2015). *Doing Bayesian data analysis* (2nd ed.). Academic Press p. 759.
Lambert, B. (2018). *A student's guide to Bayesian statistics.* London: SAGE publlications Ltd p. 498.
McElreath, R. (2016). *Statistical rethinking. A bayesian course with examples in R and stan.* Boca Raton: CRC Press p. 469.
Muth, C., Oravecz, Z., & Gabry, J. (2018). User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quantitative Methods of Psychology, 14*(2), 99–119. https://doi.org/10.20982/tqmp.14.2.p099.
Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research, 62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006.
Nuzzo, R. (2014). Statistical errors. *Nature, 506*, 150–152. https://doi.org/10.1136/bmj.1.6053.66.
Rial-Otero, R., Simal-Gándara, J., Pose-Juan, E., López-Fernández, O., & Yáñez, R. (2018). Modelling the isothermal degradation kinetics of metrafenone and mepanipyrim in a grape juice analog. *Food Research International, 108*(March), 339–346. https://doi.org/10.1016/j.foodres.2018.03.058.
Stewart, W., & Caracotsios, M. (2008). *Computer-aided modeling of reactive systems.* Hoboken, NJ: Wiley-Interscience p. 267.
Stewart, W., Caracotsios, M., & Sørensen, J. (1992). Parameter estimation from multi-response data. *AIChE Journal, 38*(5), 641–650. https://doi.org/10.1002/aic.690380502.
Tellinghuisen, J. (2015). Using least squares for error propagation. *Journal of Chemical Education, 92*(5), 864–870. https://doi.org/10.1021/ed500888r.
Van Boekel, M. (1996). Statistical aspects of kinetic modeling for food science problems. *Journal of Food Science, 61*(477–485), 489.
Van Boekel, M. (2008). *Kinetic modeling of reactions in foods.* Boca Raton: CRC/Taylor & Francis.
Zhang, L., Chen, X. D., Boom, R. M., & Schutyser, M. A. (2017). Thermal inactivation kinetics of β-galactosidase during bread baking. *Food Chemistry, 225*, 107–113. https://doi.org/10.1016/j.foodchem.2017.01.010.
Zwinderman, T. C. A. (2018). *Modern Bayesian statistics in clinical research.* Springer p. 188.